Molecular bases behind the Xg(a−) phenotype: from disruption of GATA1-regulated transcription to gene deletion

Jill R. Storry, PhD

Hematology and Transfusion Medicine, Lund University

On behalf of:

Mattias Möller, Yan Quan Lee, Karina Vidovic, Sven Kjellström, Linda Björkman, Greg Halverson, Vanya Karamatic Crew, Nicole Thornton, Martin L. Olsson

British Blood Transfusion Society | #BBTS2019

Disclosures

None

Xg^a was the first blood group assigned to a chromosome

- Discovered by Mann *et al*. (Lancet, 1962)
- Skewed frequencies between genders
  - ✓ ~30% of men are Xg(a–)
  - ✓ ~10% of women are Xg(a–)
- Xg protein is lacking on RBCs of those who are Xg(a–)
- XG escapes X-inactivation (1st gene shown)

Phenotypic relationship between Xg^a and CD99

- CD99 is the 2nd antigen in the XG system
  - ✓ ~100% of all people are CD99 positive

	Xg^a type	CD99 level
Male	Xg(a+)	High
	Xg(a–)	High or low
Female	Xg(a+)	High
	Xg(a+^{w})	High
	Xg(a–)	Low

Reid ME, Lomas-Francis C, Olsson ML. The Blood Group Antigen FactsBook, 3rd ed. 2012

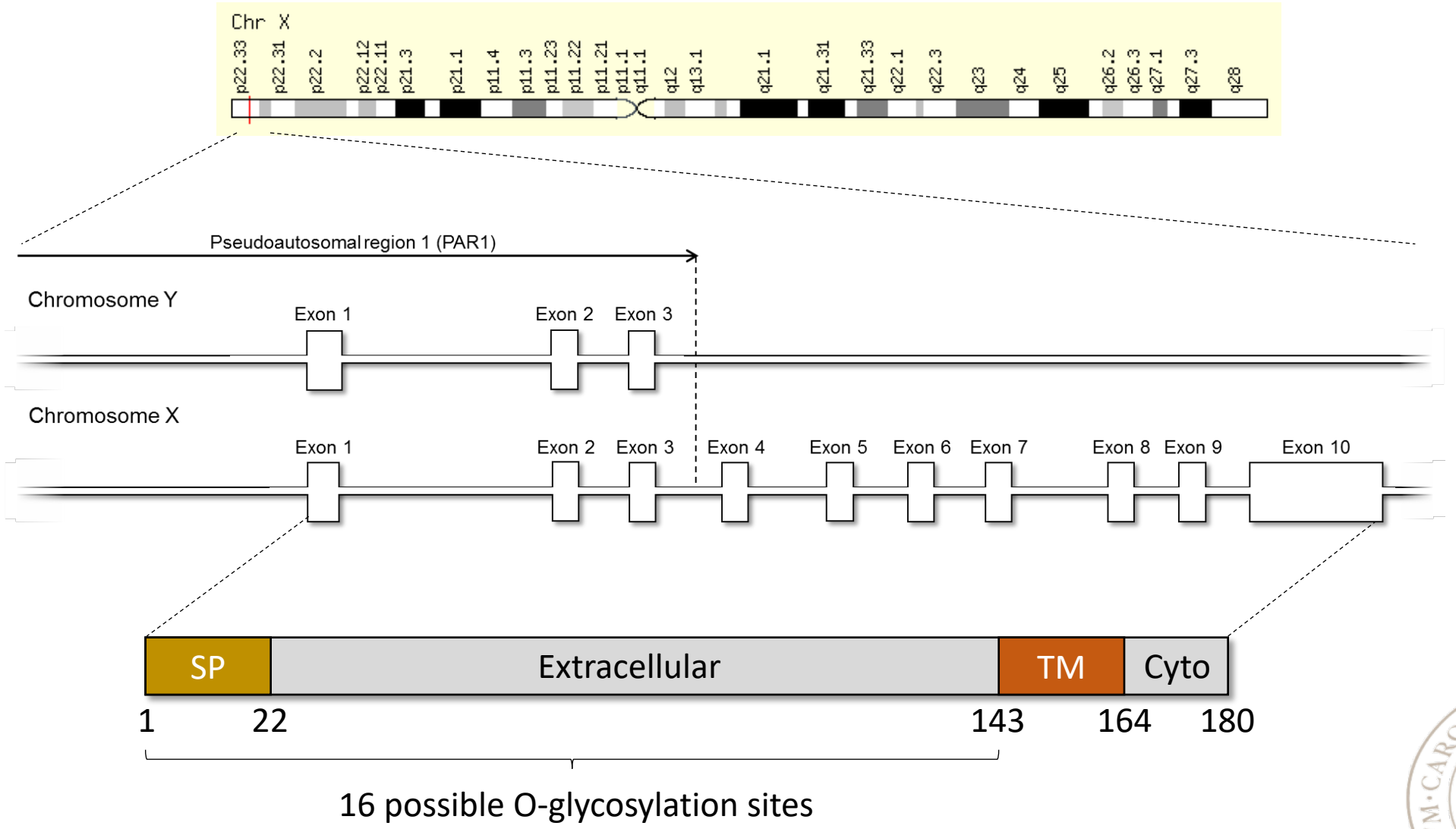
Genetic findings

- The *PBDX* gene was identified to encode Xg glycoprotein (Ellis et al. Nat Genet 1994)
- However, no explanation for presence/absence of Xg^a
- CD99 is encoded by the *MIC2/CD99* gene
- Rare CD99-negative individuals have different deletions in the coding regions of *MIC2* (Thornton et al. Vox Sang. 2015)
- A hypothetical regulatory site, *XGR*, was proposed already in 1987 (Goodfellow et al. Ann Hum Genet. 1987)

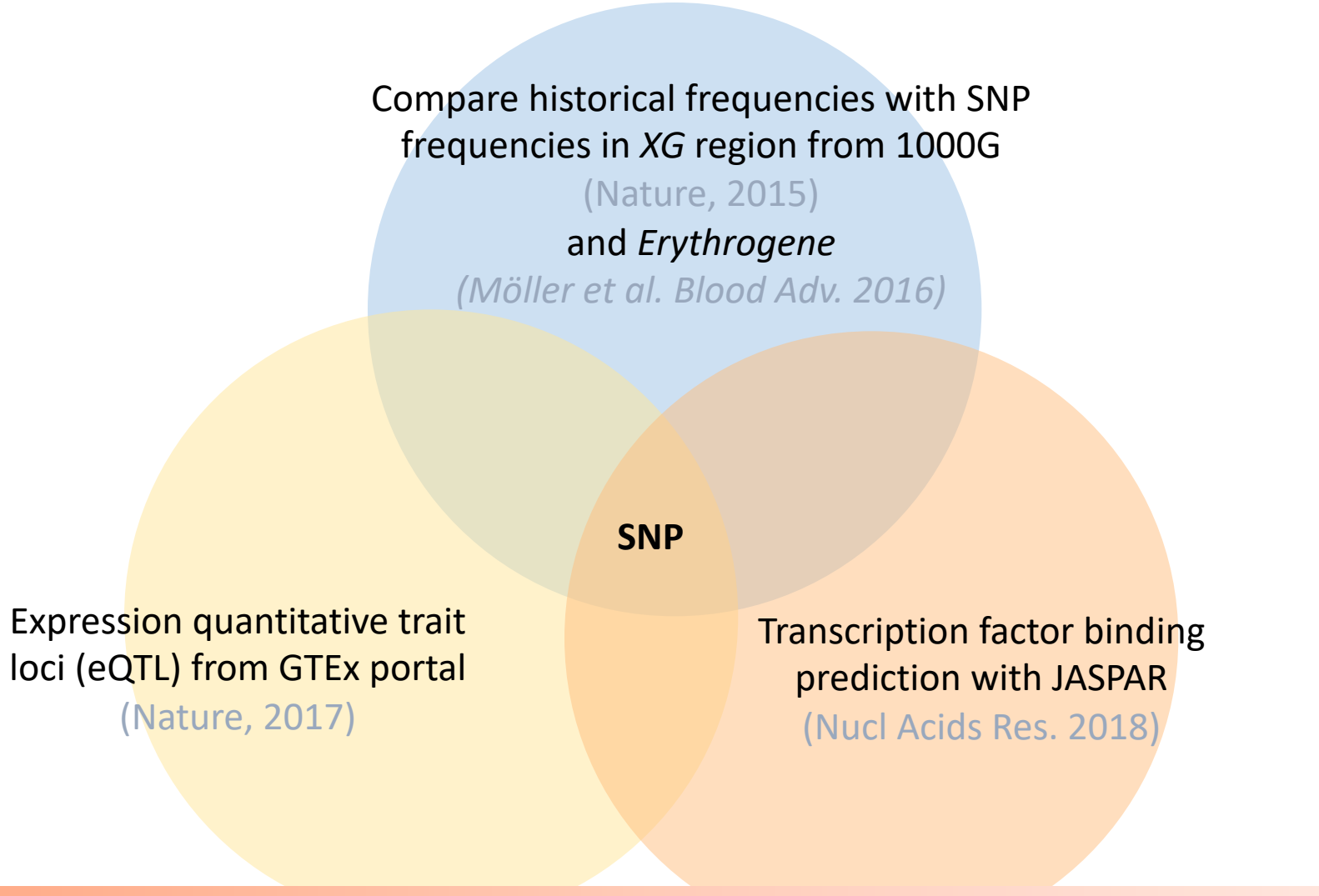
Our hypothesis:

Xg^a expression is transcriptionally regulated by a single SNP within the *XG* region, potentially disrupting an erythroid transcription factor binding site

The *XG* gene and its product



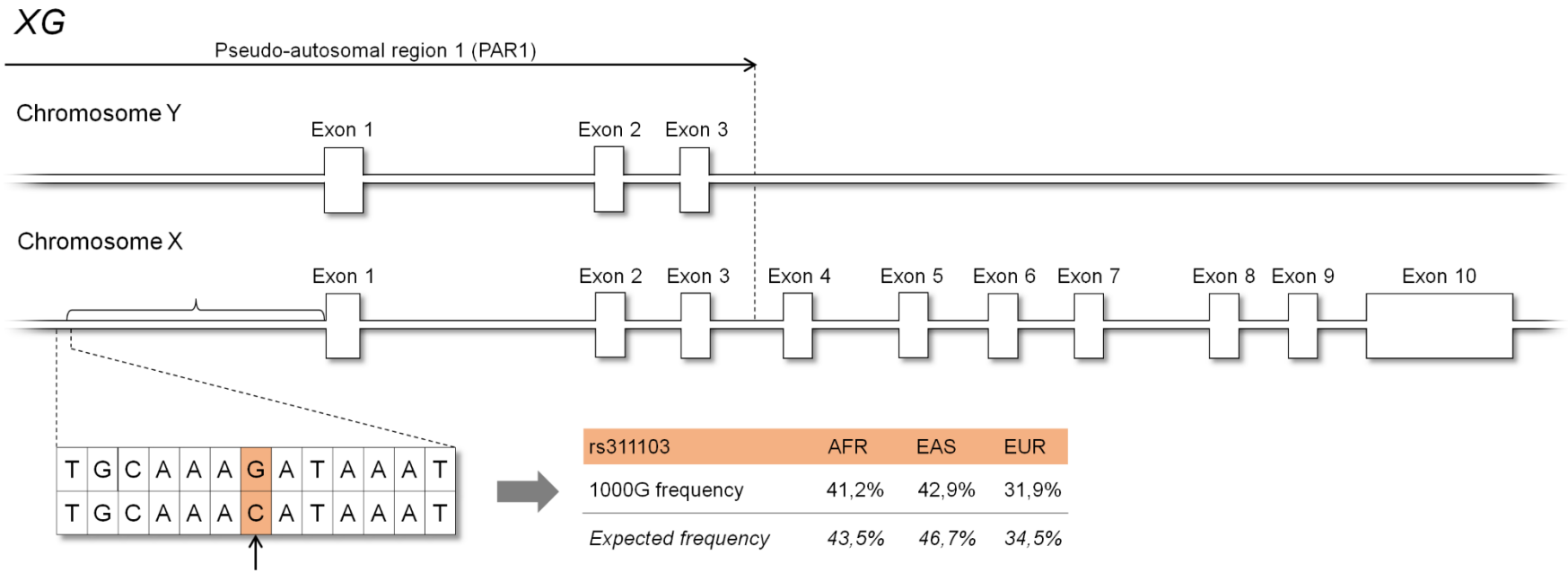
Three-pronged bioinformatics strategy



Blood samples from 158 blood donors anonymized other than for gender: Xg^a phenotyping, FACS, qPCR, EMSA, luciferase etc

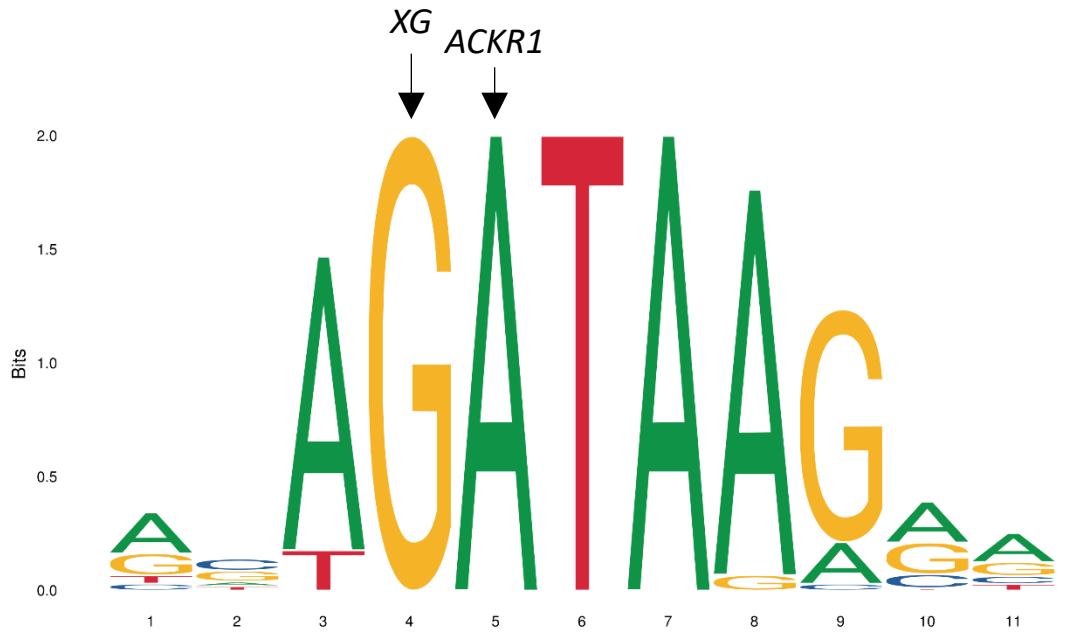
A SNP upstream of *XG* correlates with the expected phenotype distribution

Among **2,612** investigated genetic variants in the *XG* region, one specific SNP (rs311103), ~4 kb upstream of the transcription start site, was identified to have the strongest correlation to the expected distribution.

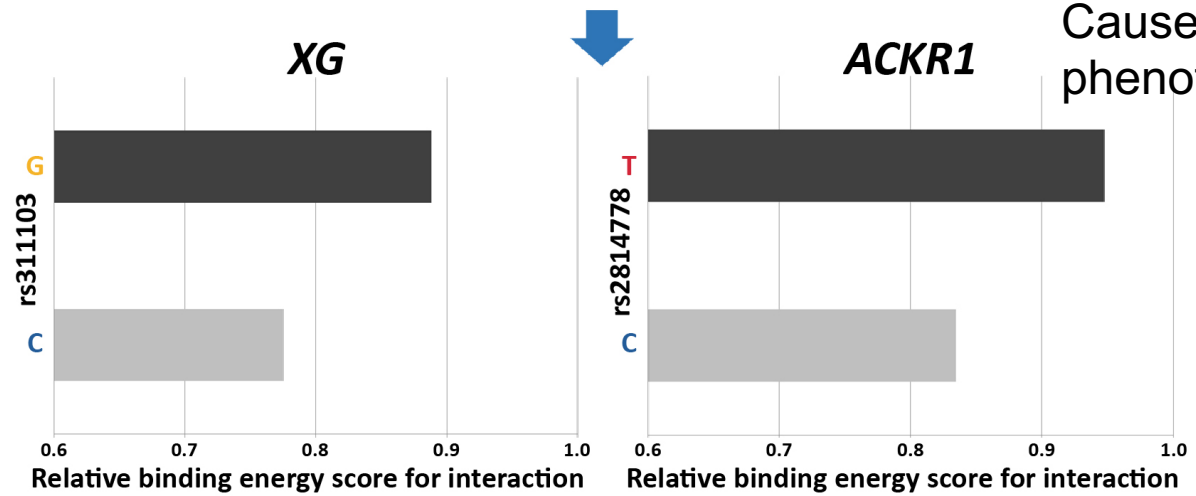


rs311103	AFR	EAS	EUR
1000G frequency	41,2%	42,9%	31,9%
Expected frequency	*43,5%*	*46,7%*	*34,5%*

The implicated SNP abolishes a potential GATA motif



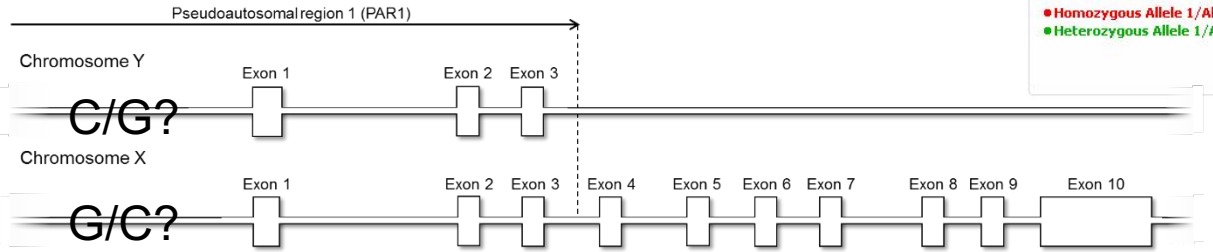
GATA1 binding (murine)

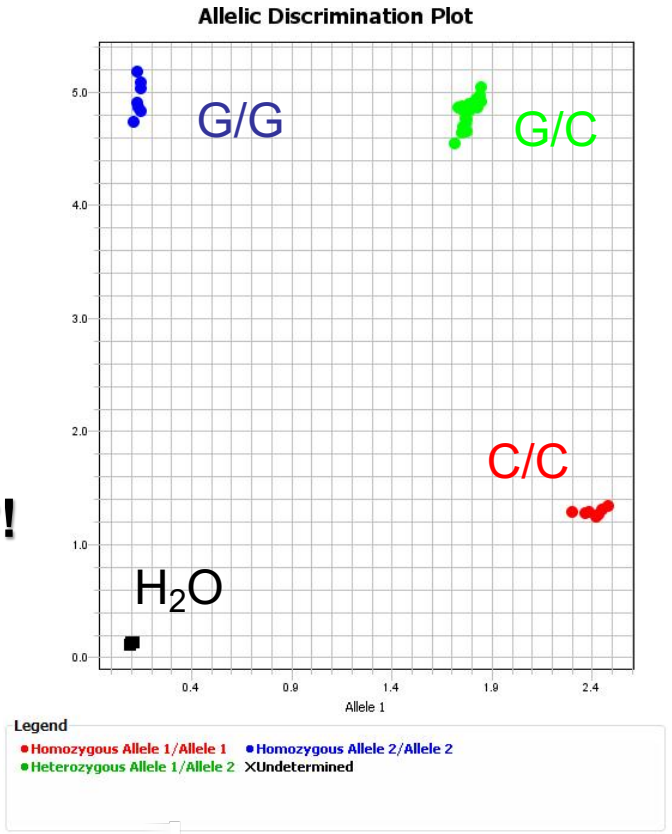


Causes Fy(a–b–) phenotype in Africans

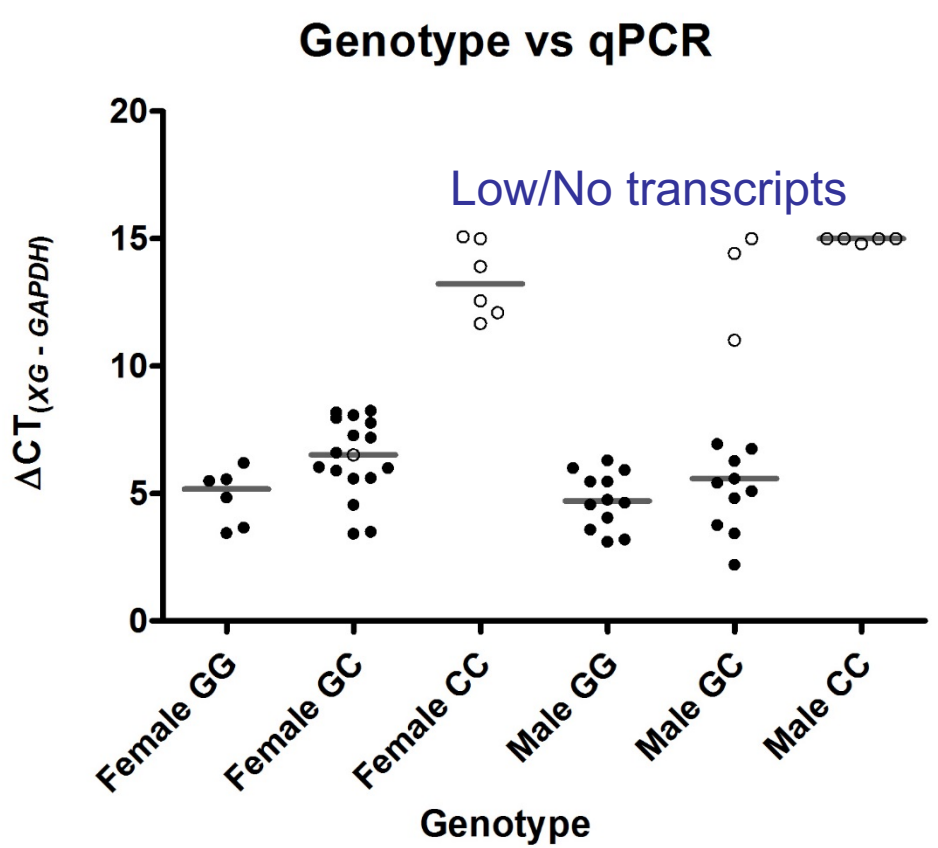
SNP genotyping by allelic discrimination

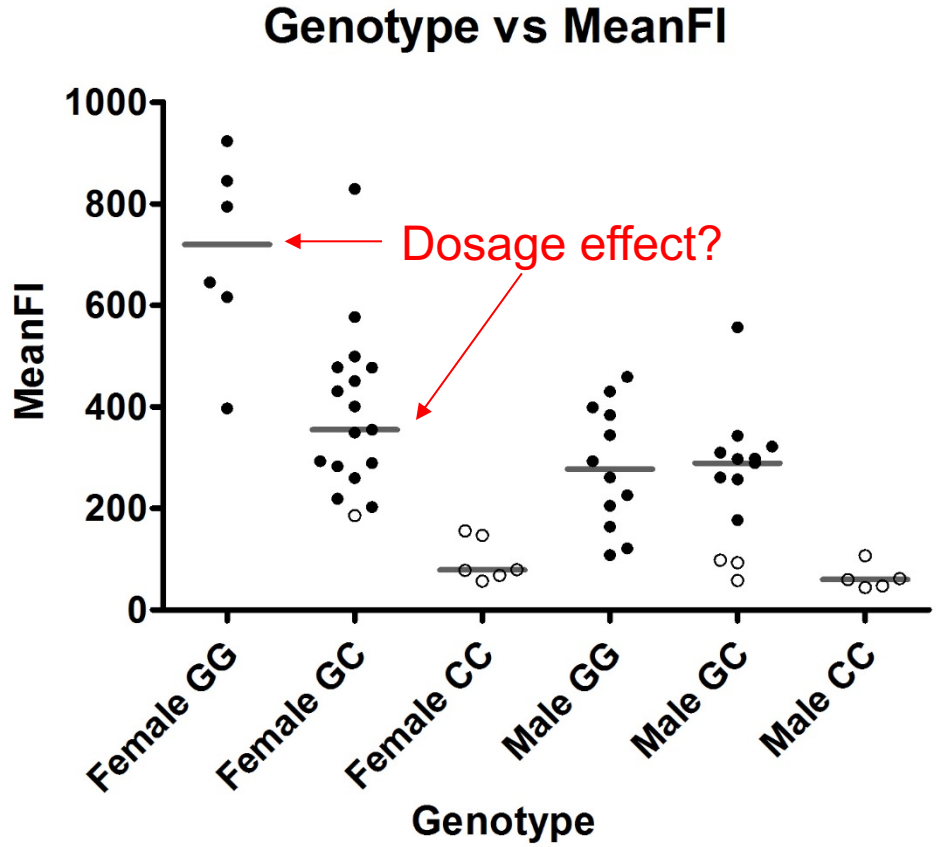
	Serology			
	Female		Male	
Genotype	Xg(a+)	Xg(a−)	Xg(a+)	Xg(a−)
GG	29	0	32	0
GC	31	1*	28	11 ?!
CC	0	13	0	13





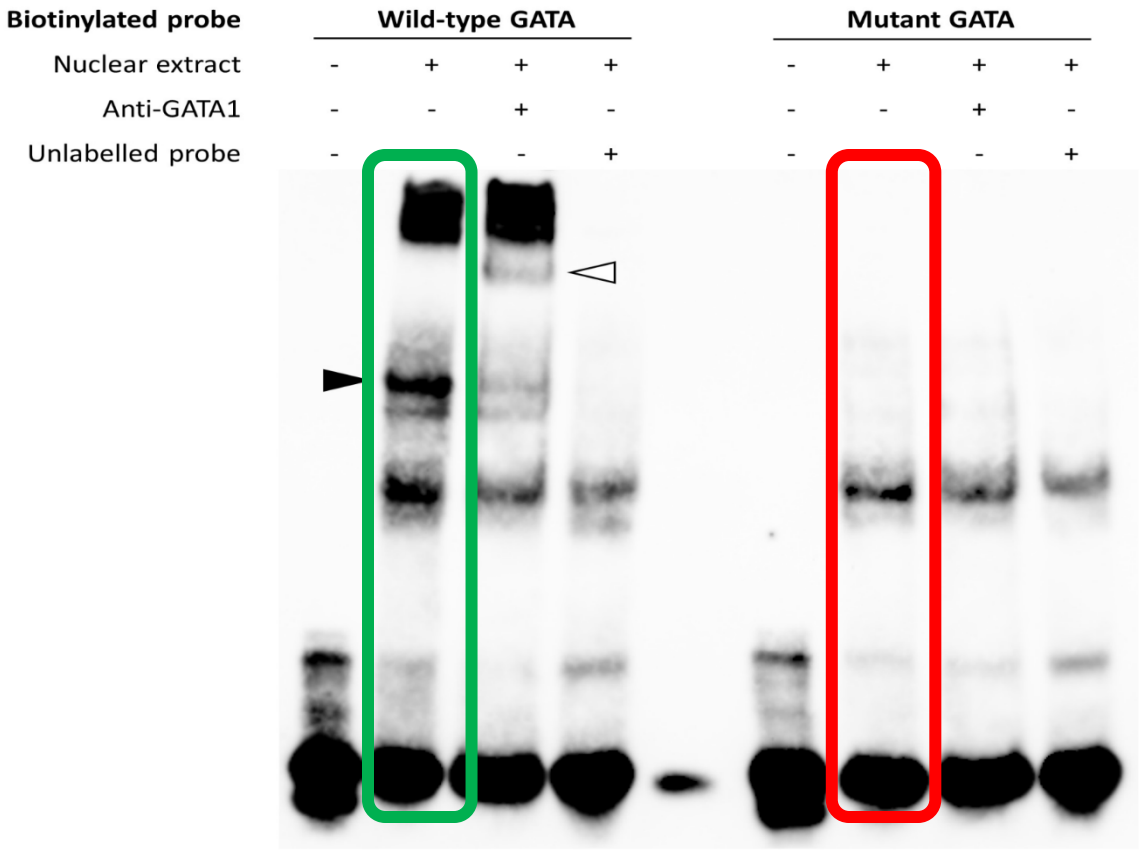
This GATA disruption abolishes *XG* transcripts
Wildtype-homozygous women are Xg^a strong



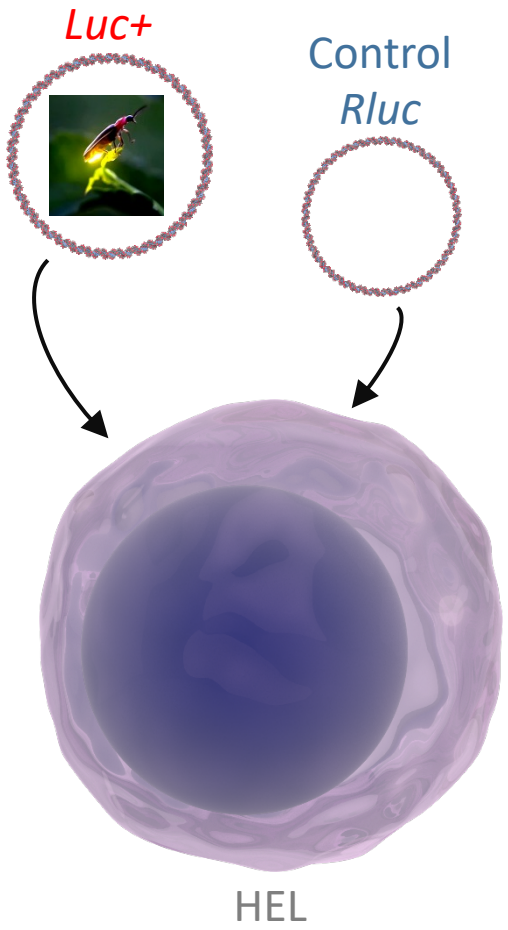


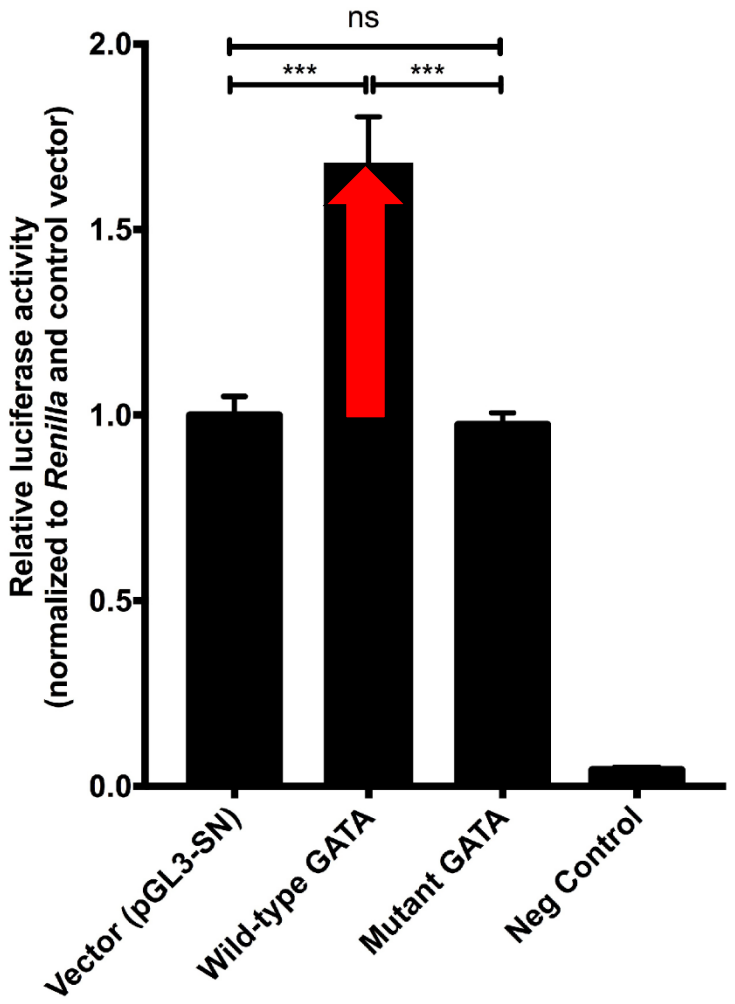
● = Xg(a+)
○ = Xg(a–)

EMSA shifts and supershifts indicate that GATA1 binds to wild-type but not mutant motif

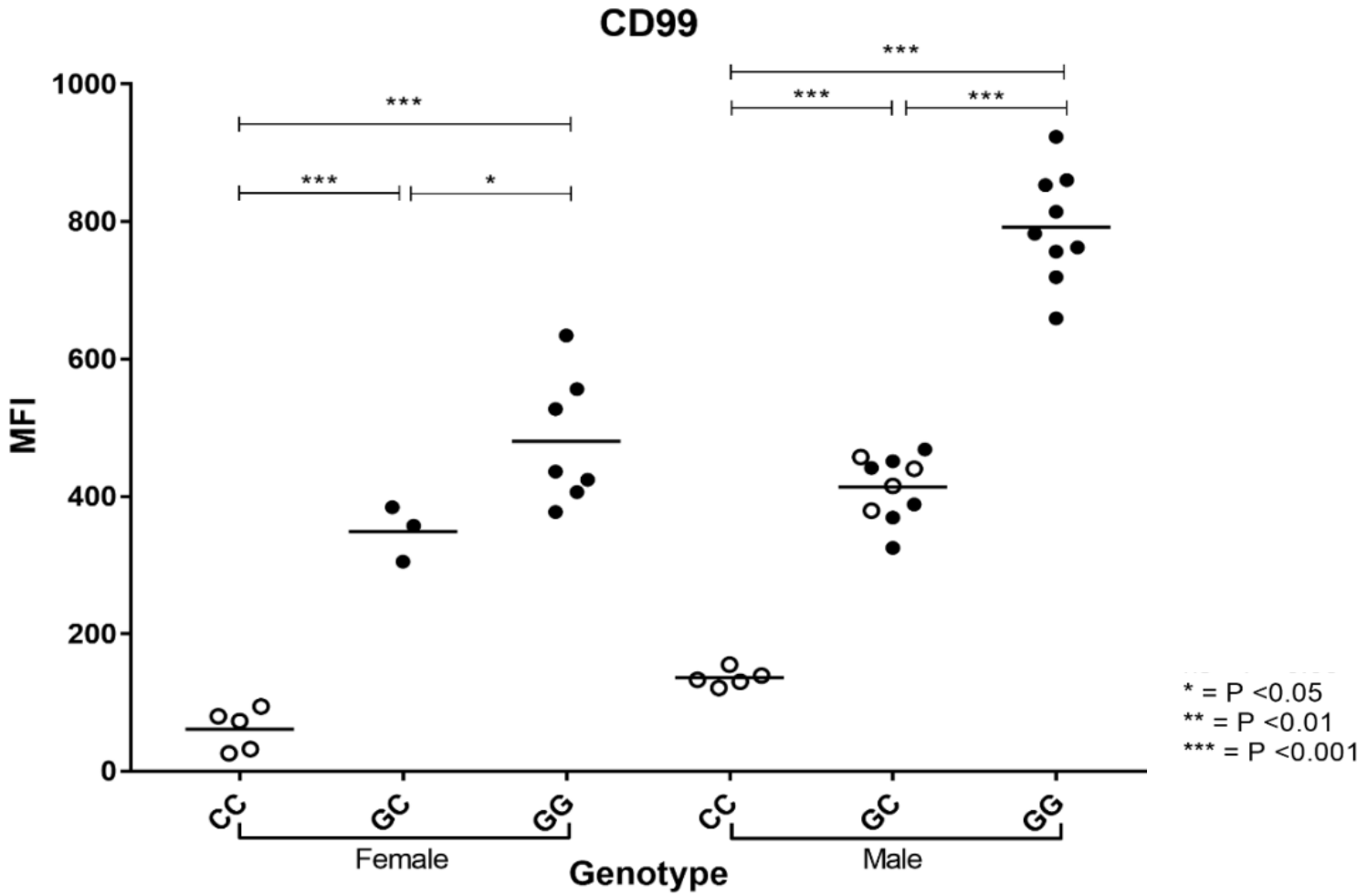


A luciferase reporter assay shows this GATA binding site to exert clear enhancer effects





rs311103 genotype also correlated well with CD99 expression level by FACS



Marion Darlison. M.Sc. Thesis, Lund University 2018

Conclusions part 1

- We could explain why a third of all men and 10% of all women lack the Xg protein on their RBCs.
- Genotyping for rs311103 predicts Xg^a status and correlates with CD99 expression levels.
- Challenges in “G/C” males (X/Y) need to be addressed.
- But variant is erythroid-specific so cannot explain why some Xg(a−) make antibodies

Disruption of a GATA1-binding motif upstream of *XG*/*PBDX* abolishes Xg^{a} expression and resolves the Xg blood group system

Mattias Möller,[1] Yan Quan Lee,[1] Karina Vidovic,[1] Sven Kjellström,[2,3] Linda Björkman,[4] Jill R. Storry,[1,4] and Martin L. Olsson[1,4]

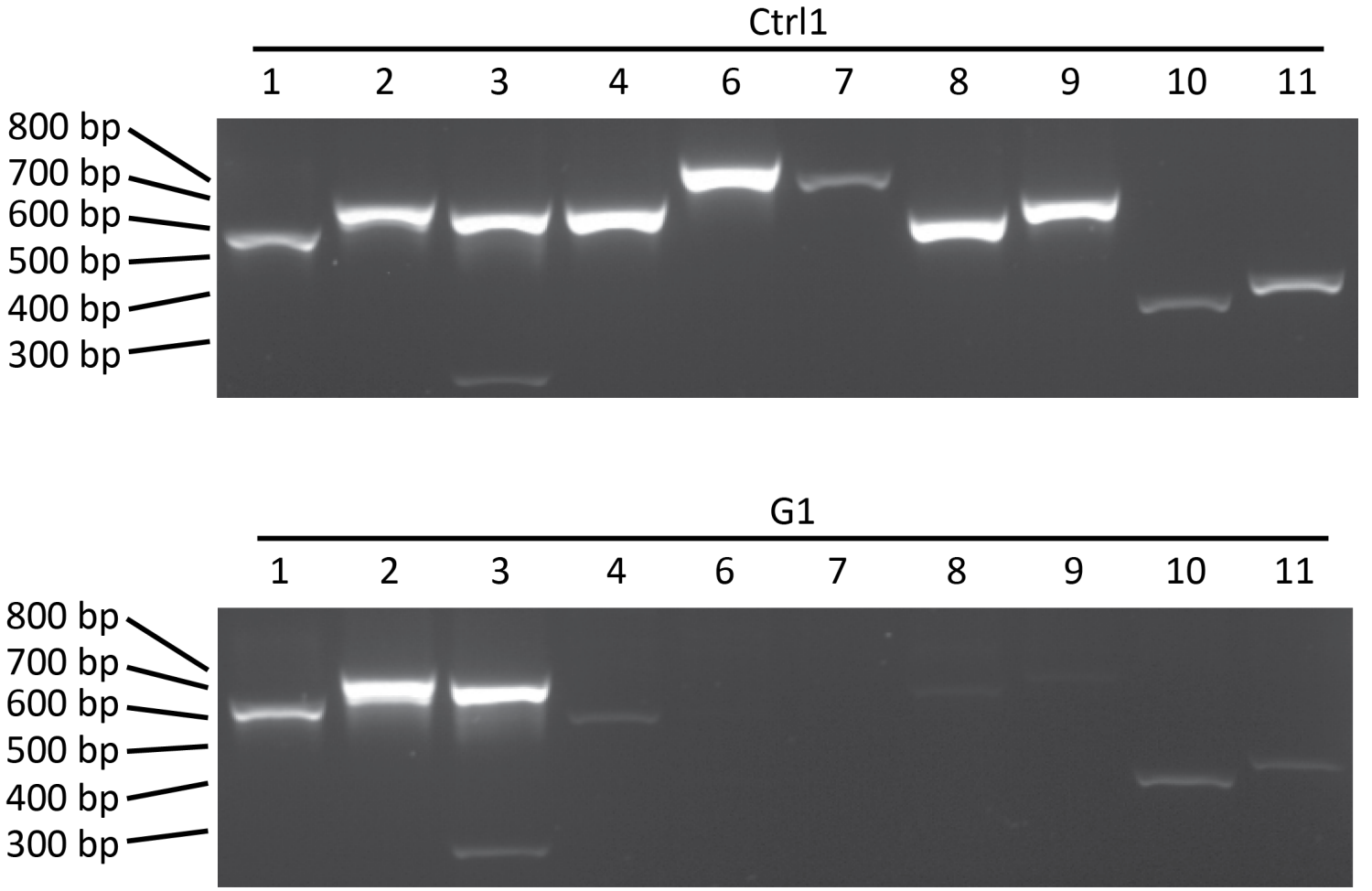
Blood. **132**, 334–338 (2018).

Time for cake!

G>C in GATA site cannot explain anti-Xg^a

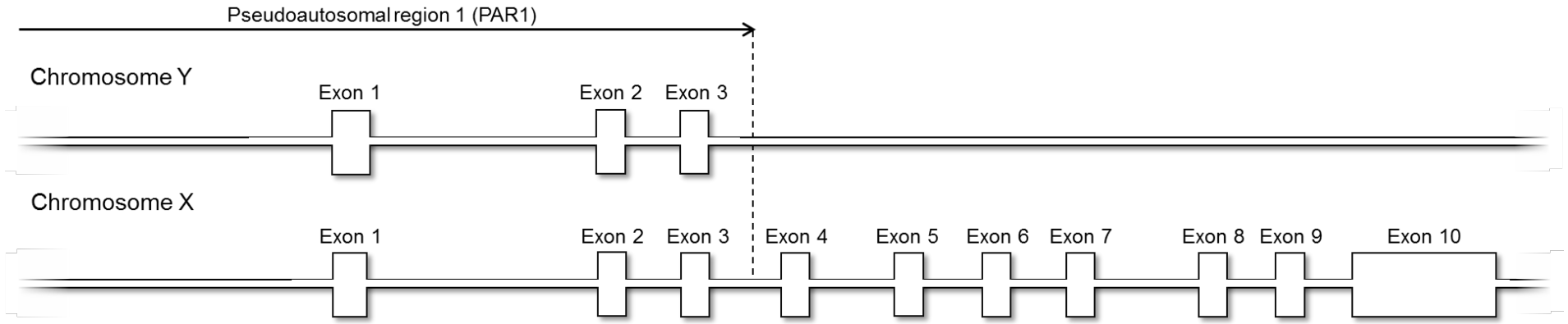
- Xg(a−) phenotype is common, but anti-Xg^a makers are exceedingly rare
- GATA1 is an erythroid-specific transcription factor; the Xg protein may be expressed elsewhere
- Managed to obtain gDNA of one male anti-Xg^a maker (G1)
  - ✓ Plan: sequence the exons of *XG* and look for non-synonymous mutations

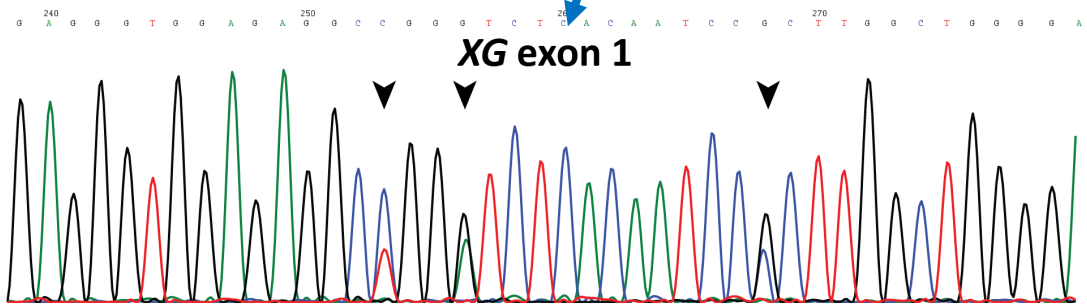
PCR for exons: missing 4-11

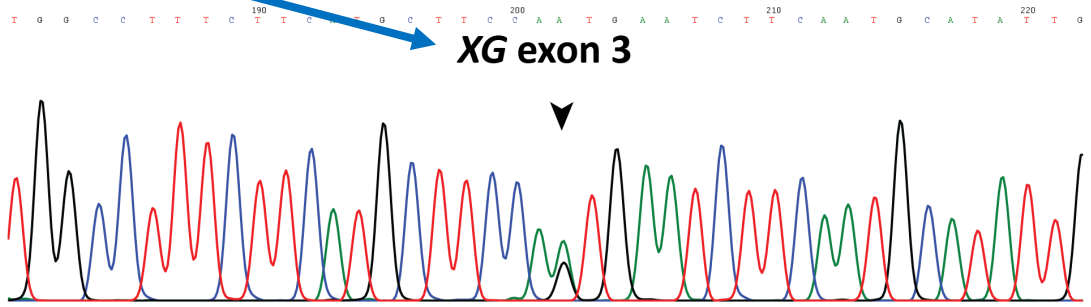


(11 exons in isoform 3)

Exons 1, 3 of *XG* are intact



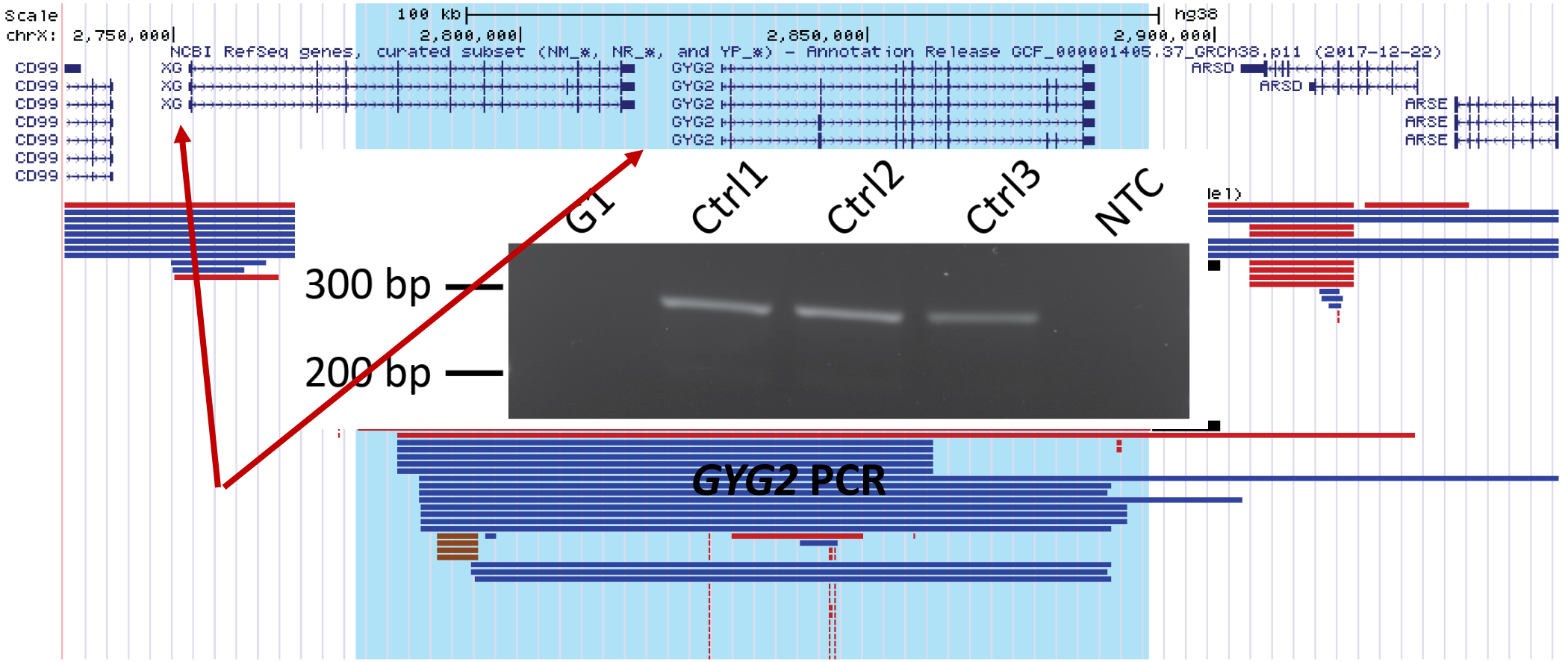




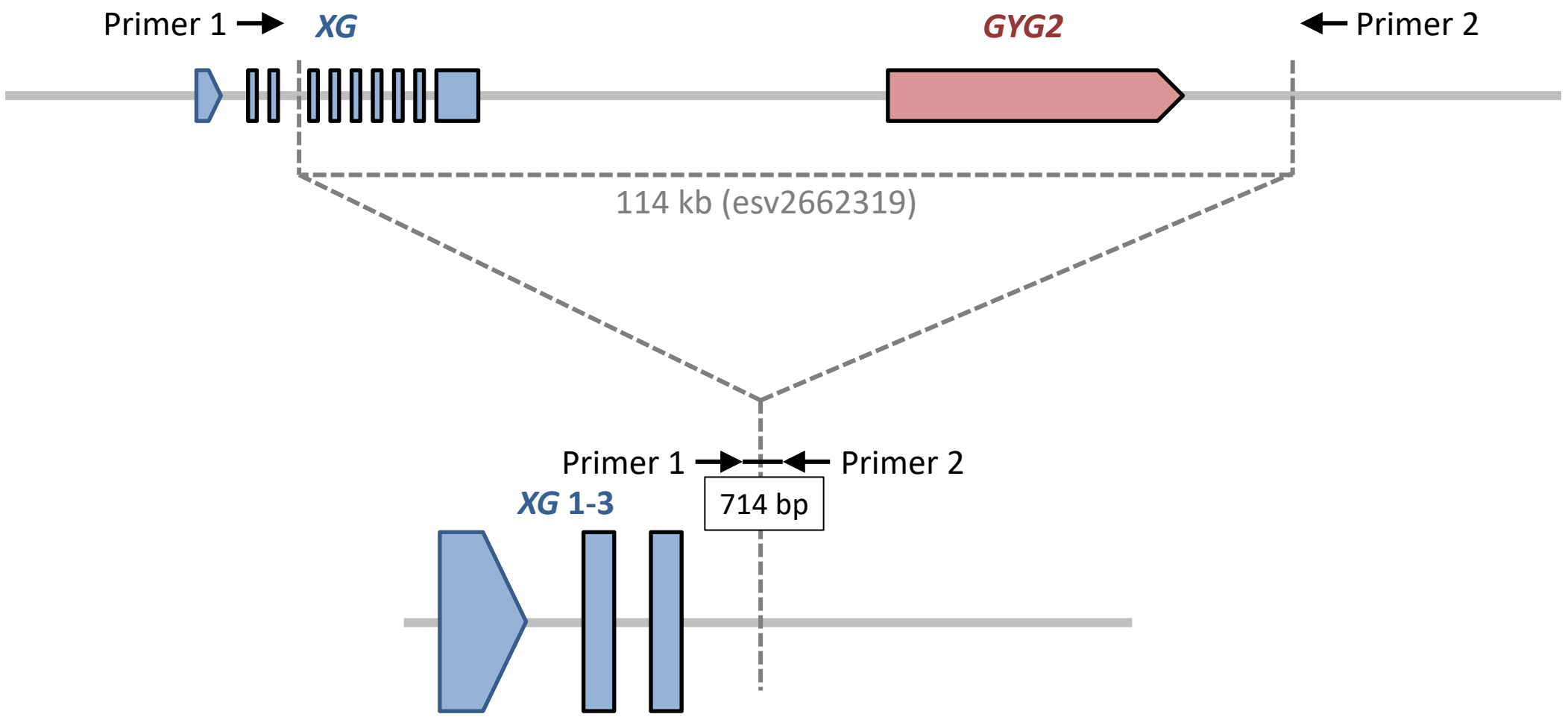
Most likely cause: esv2662319

Database of Structural Variants

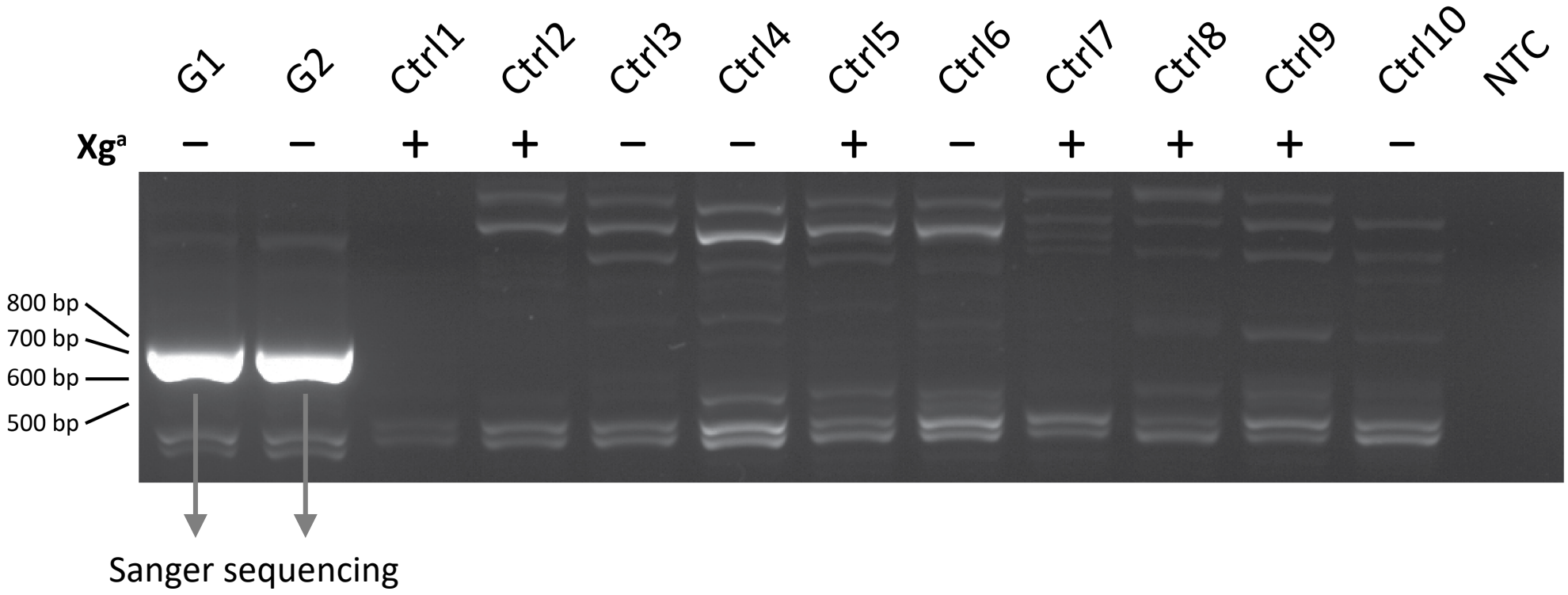
Structural variant called in 23 samples in Phase 1 of the 1000 Genomes Project



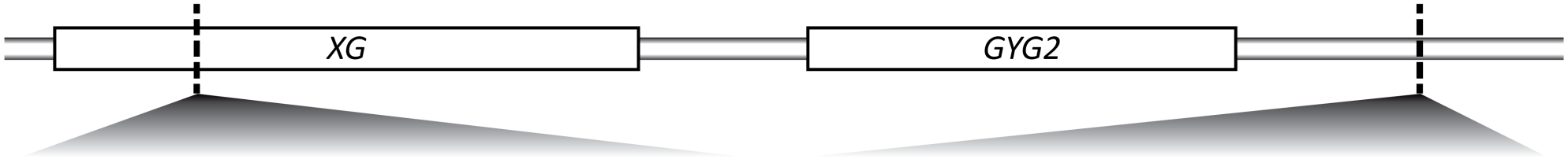
PCR to detect esv2662319: short amplicon



Validation in gDNA samples



BLAST alignment of 714 bp amplicon



Range 1: 2776038 to 2776552 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
917 bits(496)	0.0	513/520(99%)	5/520(0%)	Plus/Plus

Features: glycoprotein Xg isoform X3
glycoprotein Xg isoform 2 precursor

Forward sequence
Upstream LTR6B

```
Query  1        CCAGCACTTTGGGAGGCCAAGGCGGGTGGATCACAAGGTCAGGAGATCGAGACCATCCTG  60
Sbjct  2776038  CCAGCACTTTGGGAGGCCAAGGCGGGTGGATCACAAGGTCAGGAGATCGAGACCATCCTG  2776097

Query  61       GCTAACACGGTATGTTGTACCTGAGCGAGTTAGAAAAACGCCACACTTTGAGACAAATTA  120
Sbjct  2776098  GCTAACACGGTATGTTGTACCTGAGCGAGTTAGAAAAACG-----CTTTGAGACAAATTA  2776152

Query  121      AGAGTCCTTTATAAGCCAGCGACCGAGAGACGGCTAATGCTTAATATTCTCTCGGTCCTG  180
Sbjct  2776153  AGAGTCCTTTATAAGCCAGCGACCGAGAGACGGCTAATGCTTAATAGTCTCTCGGTCCTG  2776212

Query  181      AGGAAGGGGCTTGATTAACTTTTAGATCTTGGTTTAGGAAGGGGAGGGCTGGGGGTCTAG  240
Sbjct  2776213  AGGAAGGGGCTTGATTAACTTTTAGATCTTGGTTTAGGAAGGGGAGGGCTGGGGGTCTAG  2776272

Query  241      TGAAAACCATTTTACAGAAGTAAAGTAGGCAAAAAGTTAAAAGGATAAATGGTTGCAGGA  300
Sbjct  2776273  TGAAAACCATTTTACAGAAGTAAAGTAGGCAAAAAGTTAAAAGGATAAATGGTTGCAGGA  2776332

Query  301      AAGTAAACAGTTCCAGGTGCAGGGGCTTTAAGACTATTACAAGGTGATAGACGCGAGGCT  360
Sbjct  2776333  AAGTAAACAGTTCCAGGTGCAGGGGCTTTAAGACTATTACAAGGTGATAGACGCGGGGCT  2776392

Query  361      TTGGGCGTTACTAATCAGACGAATTCCCGGGAACTGCGGATGTAGCTCGCCACAGTATCT  420
Sbjct  2776393  TTGGGCGTTACTAATCAGACGAATTCCCGGGAACTGCGGATGTAGCTCGCCACAGTATCT  2776452

Query  421      TATCAGTTAACTGCATTCTTGGATGTGCTGGGAGTCAGCCTGCACGAGTTCAGTCCTTGA  480
Sbjct  2776453  TATCAGTTAACTGCATTCTTGGATGTGCTGGGAGTCAGCCTGCACGAGTTCAGTCCTTGA  2776512

Query  481      GGAAGGGGCTGCCAGTGAAAGAGCCAAGGTGGAGTCTGGC  520
Sbjct  2776513  GGAAGGGGCTGCCAGTGAAAGAGCCAAGGTGGAGTCTGGC  2776552
```

Range 2: 2890508 to 2890947 GenBank Graphics

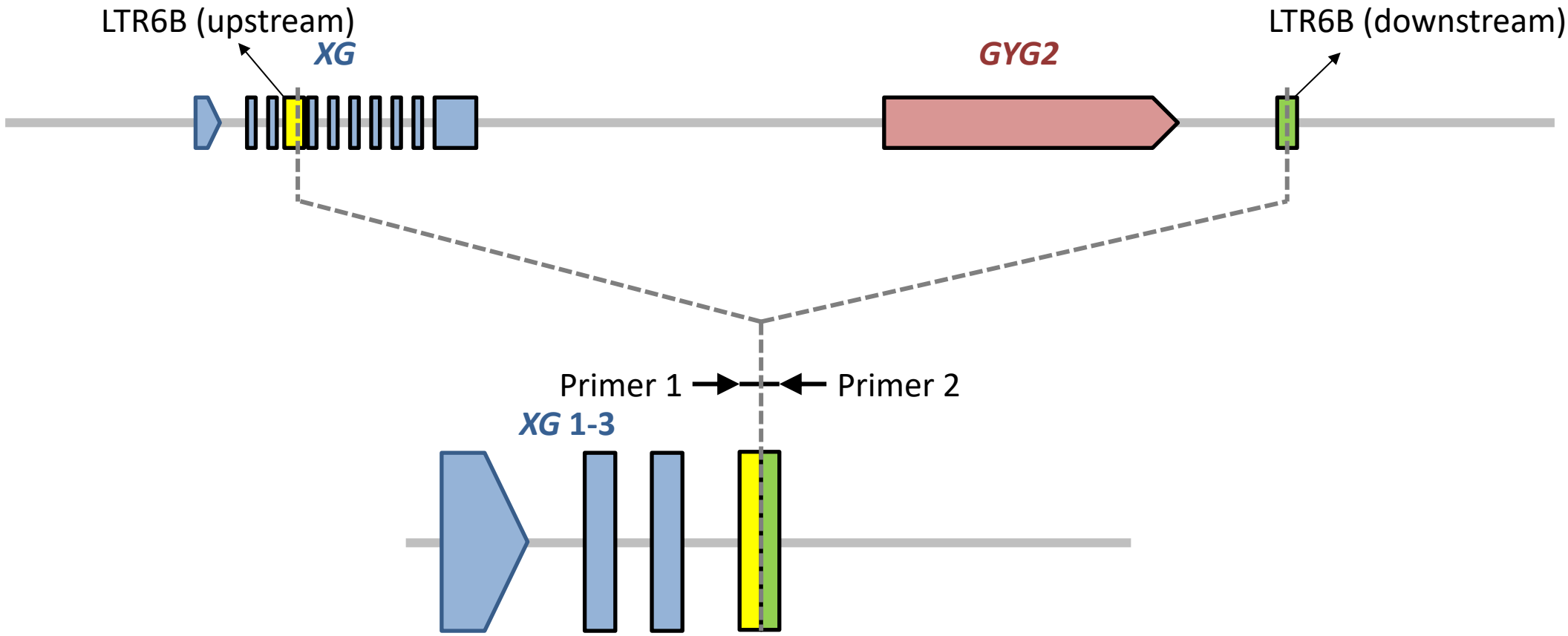
Score	Expect	Identities	Gaps	Strand
715 bits(387)	0.0	429/448(96%)	8/448(1%)	Plus/Plus

Features: 9295 bp at 5' side: glycogenin-2 isoform e
16324 bp at 3' side: arylsulfatase D isoform X1

Forward sequence
Downstream LTR6B

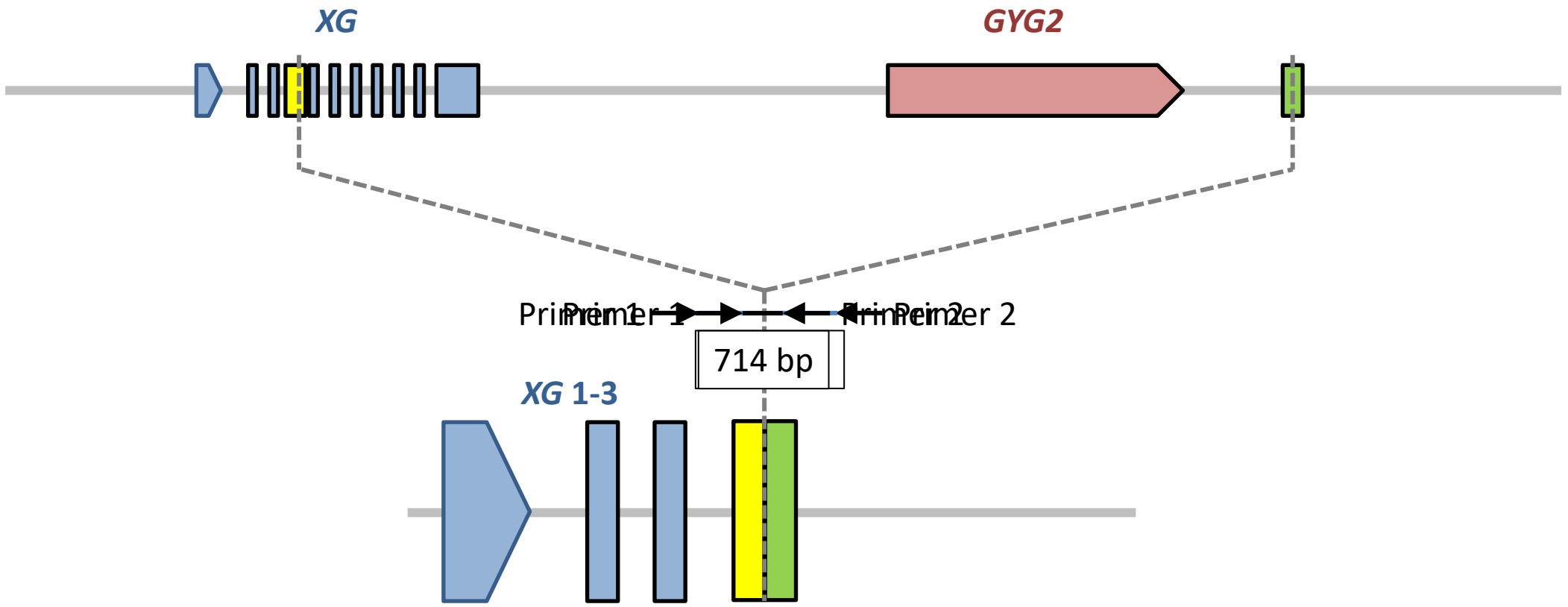
```
Query  73       TGTTGTACCTGAGCGAGTTAGAAAAACGCCACACTTTGAGACAAATTAAGAGTCCTTTAT  132
Sbjct  2890508  TGTTGTACCCGAGCGAGTTAGAAAAACGCCACACTTTGAGACGATTTAAGAGTCCTTTAT  2890567

Query  133      AAGCCAGCGACCGAGAGACGGCTAATGCTTAATATTCTCTCGGTCCTGAGGAAGGGGCTT  192
Sbjct  2890568  TAGCCGGCGACCGAGAGACGGCTAACGCTCAAAATTCTCTCGGCCCCGAGGAAGGGGCTT  2890627

Query  193      GATTAACTTTTAGATCTTGGTTTAGGAAGGGGAGGGCTGGGGGTCTAGTGAAAACCATTT  252
Sbjct  2890628  GATTAACTTTTAGATCTTGGTTTAGGAAGGGGAGGGCGGGGGGTCTAGTGAAAACCATTT  2890687

Query  253      TACAGAAGTAAAGTAGGCAAAAAGTTAAAAGGATAAATGGTTGCAGGAAAGTAAACAGTT  312
Sbjct  2890688  TACAGAAGTAAAGTAGGCAAAAAGTTAAAAGGATAAATGGTTGCAGGAAAGTAAACAGTT  2890747

Query  313      CCAGGTGCAGGGGCTTTAAGACTATTACAAGGTGATAGACGCGAGGCTTTGGGCGTTACT  372
Sbjct  2890748  CCAGGTGCAGGGGCTTTAAGACTATT--------ATAGACGCGAGGCTTTGGGCGTTACT  2890799

Query  373      AATCAGACGAATTCCCGGGAACTGCGGATGTAGCTCGCCACAGTATCTTATCAGTTAACT  432
Sbjct  2890800  AATCAGACGAATTCCCGGGAACTGCGGATGTAGCTCGCCACAGTATCTTATCAGTTAACT  2890859

Query  433      GCATTCTTGGATGTGCTGGGAGTCAGCCTGCACGAGTTCAGTCCTTGAGGAAGGGGCTGC  492
Sbjct  2890860  GCATTCTTGGATGTGCTGGGAGTCAGCCTGCACGAGTTCAGTCCTTGAGGAAGGGGCTGC  2890919

Query  493      CAGTGAAAGAGCCAAGGTGGAGTCTGGC  520
Sbjct  2890920  CAGTGAAAGAGCCAAGGTGGAGTCTGGC  2890947
```

Predicted breakpoints are in LTR6B motifs



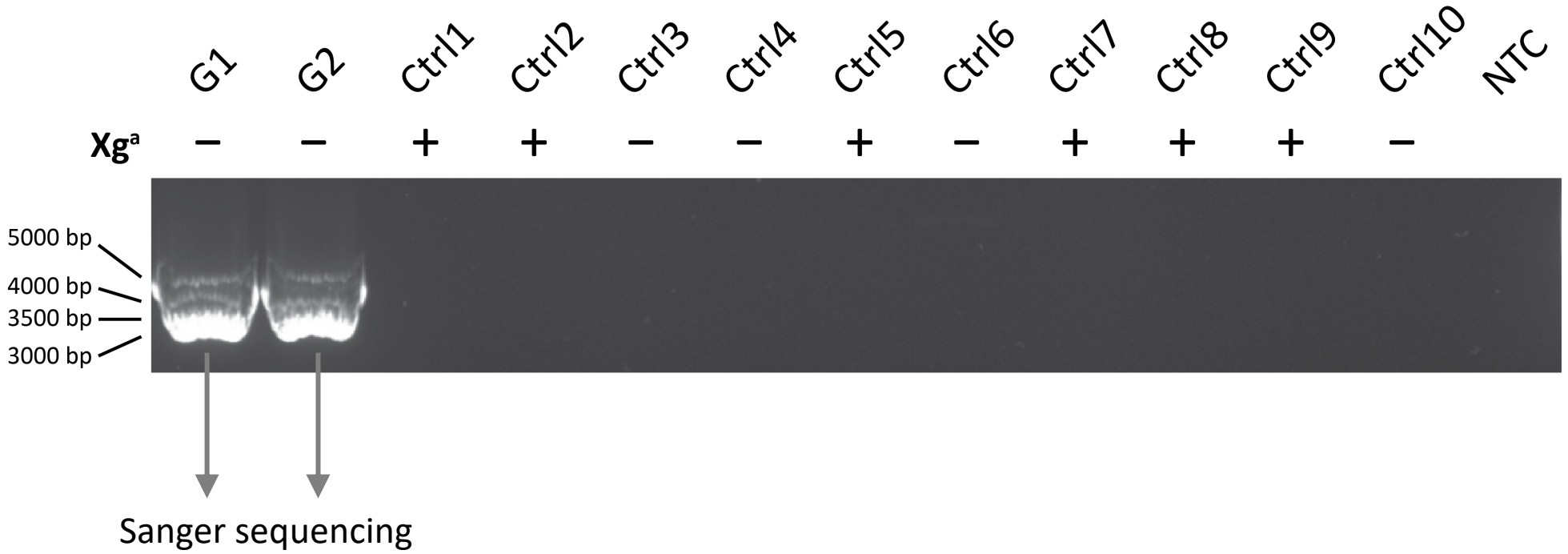
- **L**ong **T**erminal **R**epeat **6B**: HERVS71 endogenous retrovirus (>500 copies in genome)
- Sequencing of PCR amplicons: could not distinguish between upstream and downstream LTR6B sequences

PCR to detect esv2662319: long amplicon



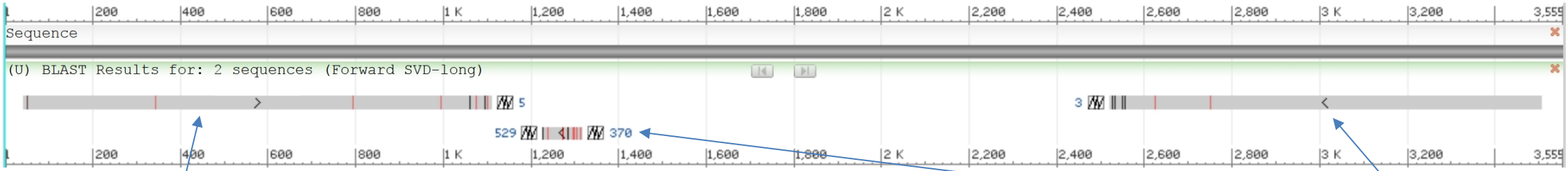
- Larger amplicon to avoid sequencing homologous regions

Validation in gDNA samples



Sequence alignment to predicted amplicon

Predicted amplicon: 3555 bp



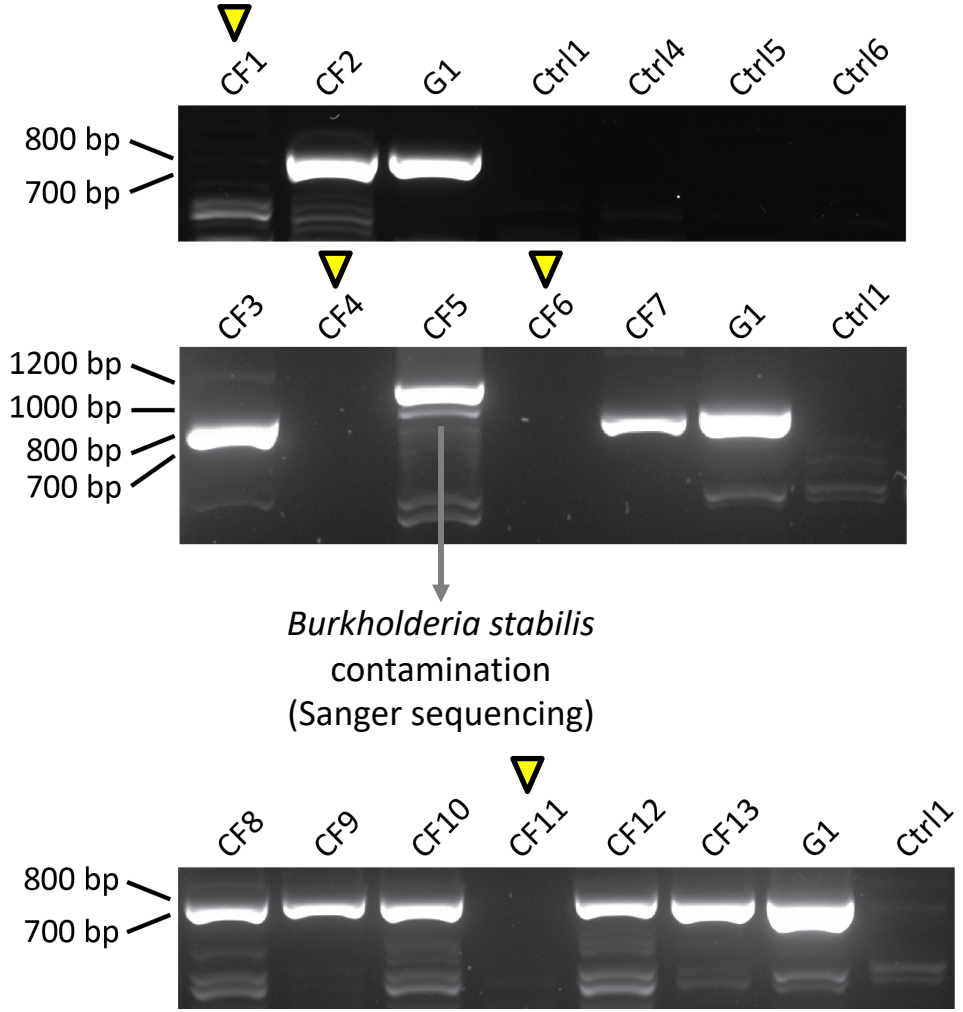
Forward sequence

Reverse sequence

Plasma samples: cell-free DNA (cfDNA)

- Obtained archived anti-Xg^a plasma samples (very old!)
- Performed cfDNA extraction from plasma
- Ran PCR for short amplicon (714 bp)

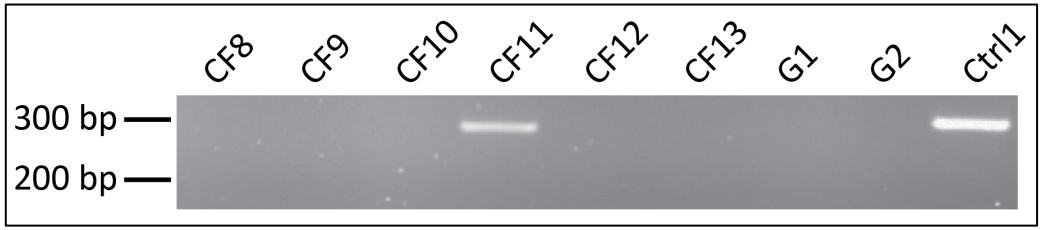
PCR with 13 cfDNA samples



- 8 of 13 positive for esv2662319
- 1 sample contaminated
- 4 samples without amplification

4 samples without amplification

- PCR for *XG* exon 4 (589 bp)
  - No bands
- PCR for *GYG2* (284 bp)
  - CF11 showed a band



- Control PCR for unrelated gene *KCNJ11* (720 bp)
  - No bands
  - cfDNA predominantly comprises short (150, 300, 450 bp) fragments

S. Volik, et al. *Mol Cancer Res*. **14**, 898–908 (2016).

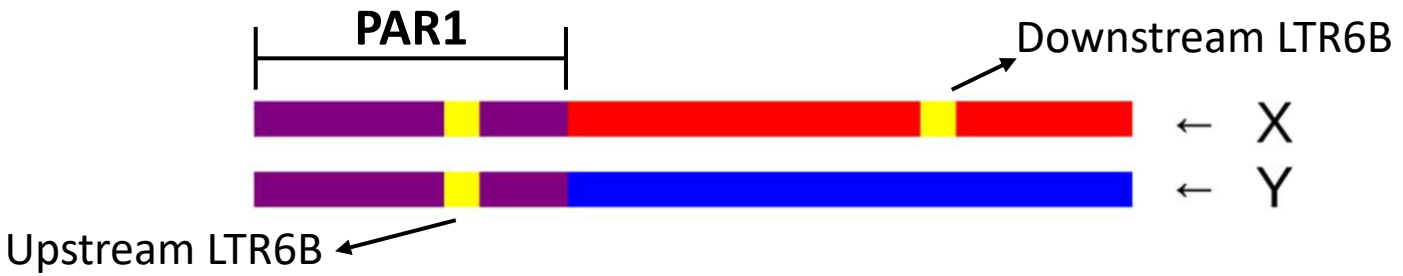
Descriptive statistics - summary

- 2 gDNA and 13 cfDNA samples
- 4 samples dropped out due to contamination or poor DNA quality (no amplification with control *KCNJ11* PCR)

- 10 of 11 remaining samples
  - ✓ positive for esv2662319 short amplicon
  - ✓ negative for *XG* exon 4
  - ✓ negative for *GYG2*

- 1 sample unexplained! (positive for *GYG2*)

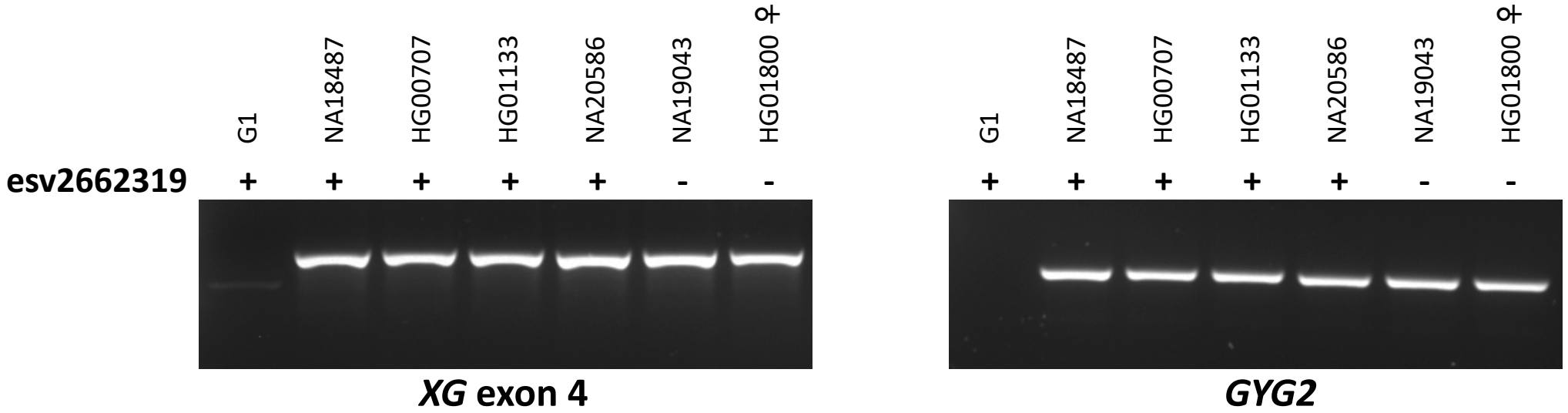
Loss of 114 kb by recombination



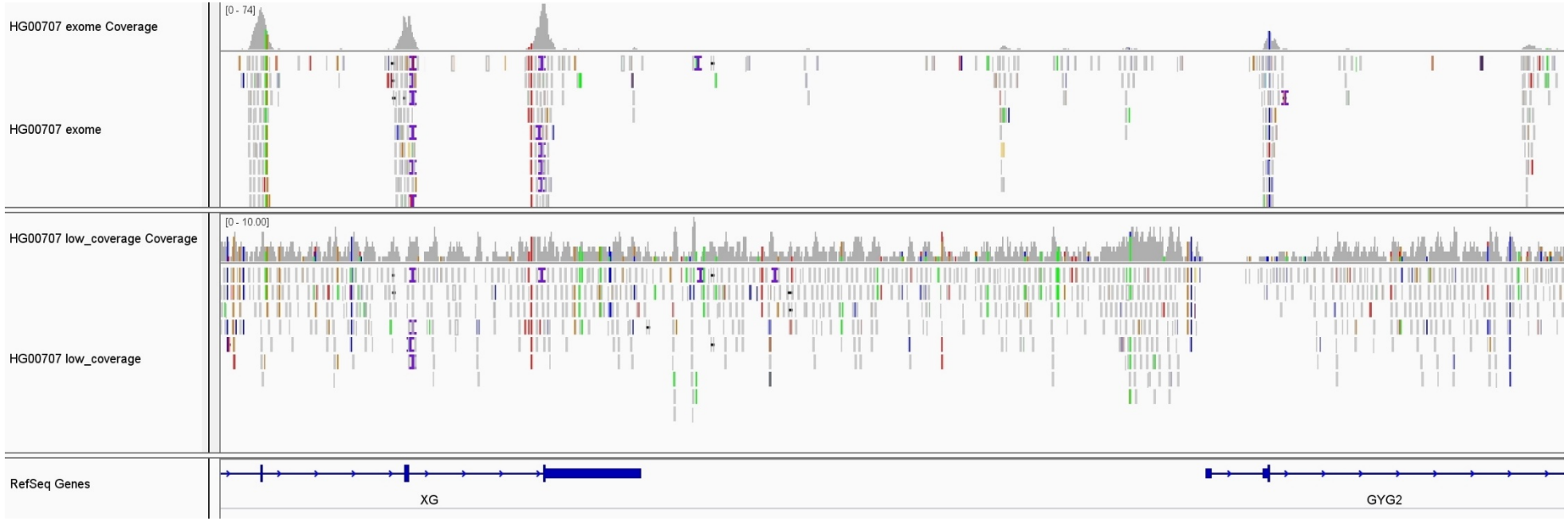
Adapted from: M. A. Mensah *et al.*, Pseudoautosomal Region 1 Length Polymorphism in the Human Population. *PLOS Genetics*. **10**, e1004578 (2014).

Bad calls in Phase 1 of 1000 Genomes Project

- Obtained gDNA from 4 male samples from 1000 Genomes project
  - ✓Structural variant esv2662319 called in these samples



Examining sequencing reads



Published May 2019

A large deletion spanning *XG* and *GYG2* constitutes a genetic basis of the Xg_{null} phenotype, underlying anti-Xg^a production

Yan Quan Lee[1], *Jill R. Storry*[1,2], *Vanja Karamatic Crew*[3], *Gregory R. Halverson*[4], *Nicole Thornton*[3], *and Martin L. Olsson*[1,2]

Transfusion. **59**, 1843–1849 (2019).

What is the function of these proteins?

- Xg and CD99 proteins are 48% homologous
  - ✓ Similar to glycophorins
  - ✓ Large N-terminal portions heavily O-glycosylated
  - ✓ A single transmembrane domain
- Xg is relatively RBC-specific whilst CD99 is expressed in many different cell types
- CD99 shown to be an adhesion molecule
  - ✓ Roles in immunology, cancer etc
- The function of the 149-aa Xg glycoprotein is unknown
  - ✓ Based on homology – may have similar role as CD99 but on RBCs

The Blood Group @ LU

Our research aims to uncover new roles of the ***red blood cell*** surface in health and disease, with a special focus on the polymorphic molecules known as ***blood groups***.

LUND UNIVERSITY

Dept. of Hematology & Transfusion Mmedicine
Department of Laboratory Medicine

REGION SKANE

Clinical Immunology & Transfusion Medicine
LabMedicine, Office of Medical Services

Linda Björkman
M.D.

Annika Hult
Ph.D., staff scientist

Karina Vidovic
Ph.D.,staff scientist,
Lab manager

Martin L Olsson
Professor, M.D., Ph.D.
Principal investigator

Magnus Jöud
M.D., Ph.D. student

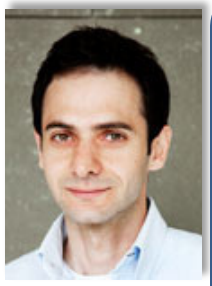
Anja Nylander
M.D., Ph.D. student

Philaiphon (Pat) Jongruamklang
M.Sc., Ph.D. student

Yan Quan Lee
Ph.D., Post-doc

Jill Storry
Assoc. Prof, Ph.D.,
Deputy group leader

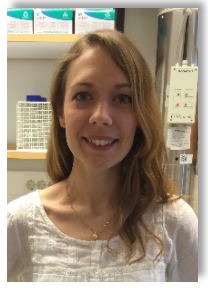
Åsa Hellberg
Ph.D., Reference laboratory coordinator



Abdul Ghani Alattar
M.D., Ph.D. student

Marion Darlison
M.Sc. student

Jennifer Ricci Hagman
M,Sc., Ph.D. student



Linn Stenfelt
M.Sc., Ph.D. student

Mattias Möller
M.D., Ph.D. student
Bioinformatician

Our work is funded by:

ALF

European Commission

Knut och Alice Wallenbergs Stiftelse

Crafoordska stiftelsen
GRUNDAD AV HOLGER CRAFOORD 1980

Vetenskapsrådet

What is *GYG2*?

- *GYG2* encodes Glycogenin 2
- Glycogenin is a self-glucosylating protein involved in blood glucose homeostasis
- Glycogen storage disease
- Neurology, failure to thrive